05: EDA for Quantitative Variables

Stat 120 | Fall 2025

Prof Amanda Luby

The lab manual data `Pew` contains data from a January 2014 Pew Research Center survey about the Internet. We'll consider three variables from this survey:

- `age`: the person's age, in years
- `income`: the person's yearly income, grouped into four categories
- `values`: an answer to the question asking if the Internet has been a "good thing"

-1. Run the code chunk below to load the libraries we need for this activity.

```
library(tidyverse)
```

0. Import the data into R by running the following code chunk. This chunk also includes a line to redefine the levels of the `income` variable so that they show up in the right order. Click on the dataset in the environment pane to pull up the data viewer. Can you see the "spreadsheet view" of the data?

```
pew = read_csv("http://www.math.carleton.edu/Stat120/RLabManual/Pew.csv")
pew$income <- factor(pew$income, levels = c("0-30000", "30000-75000",
                                            "75000-150000", "150000"))
```

1. If you haven't already, create a histogram of the `age` variable using esquisse. Make sure to copy your code to the chunk below.

2. In the empty code chunk below, run `mean(pew$age)` and `median(pew$age)`. Now try `mode(pew$age)`. What happened?

3. Compute the standard deviation of `age`. Does the 68/95/99 rule apply? Find the interval where 95% of the cases should fall.

4. Let's see if you're right! The chunk of code below `filters` the data to only include values where `age` is greater or equal to 18 and less than or equal to 99 and then spits out how many cases are in the filtered dataset. These are the minimum and maximum values, so the entire dataset is included. Edit this chunk so it corresponds to your interval values above. How close to 95% is it?

```r
pew %>%
  filter(age <= 99 & age >= 18) %>%
  summarize(
    n = n()
  )
```

```
# A tibble: 1 x 1
      n
  <int>
1   696
```

4. Are people who think the internet has been a "good thing" younger than those that don't? Create side-by-side boxplots to answer the question. (You can use ggplot, base R, or esquisse to create your boxplots)

5. Another way to look at the distribution of a quantitative variable across different levels of a categorical variable is using `facets`. Facets create a different graph for each level of a categorical variable. Choose one method below. + In esquisse, create a histogram of `age` and then drag `values` into the facet box (far right) + If using ggplot directly, create a histogram of `age` and then add `facet_wrap(vars(values))` What can you see using the histogram approach that you cannot see with the side-by-side boxplots?

6. We also often want summary statistics per group. The code below groups the data by `values`, and then computes the mean and median for each group. Add a line of code to compute the `sd()` for each group. (You may also need to add a comma at the end of a line) Do these quantities align with the visualizations that you made above?

```r
pew %>%
  group_by(values) %>%
  summarize(
    mean_age = mean(age),
    median_age = median(age)
  )
```

```
# A tibble: 2 x 3
  values mean_age median_age
  <chr>     <dbl>      <dbl>
1 bad        49.0         51
2 good       48.2         50
```

Stop here. and let Amanda know that you've finished with the `pew` data portion.

1 Part 2: Outliers and Log Transformations

For the part of the activity, we're going to be looking at the CEO salaries for S&P 500 companies. This data was gathered by AFL-CIO and more information about the data is available on their website.

0. Import the data into R by running the following code chunk. Click on the dataset in the environment pane to pull up the data viewer. Can you see the "spreadsheet view" of the data? List the variables and whether they are categorical or quantitative. Note any ID variables.

1. Using `esquisse`, `ggplot`, or `base R`, create a histogram of the `Pay` variable

2. Compute the `mean`,`median`, and `sd` of the `Pay` variable and describe the distribution. (You should focus on the *shape*, *center*, and *spread*.).

3. You should have noted that the distribution is *right-skewed* (it has a long right tail). The code below demonstrates one way to remove outliers by `filter`ing the very large observations and saving it to a new dataset called `ceo_salaries_outliers`. Run the code below, and then view the new dataset. Which companies have the highest CEO pay?

```
ceo_salaries_outliers = ceo_salaries %>%
  filter(Pay > 100000000)
```

4. Next, create a dataset called `ceo_salaries_filtered` that includes all cases where Pay is less than $100,000,000. Create a histogram of the `Pay` variable for this filtered dataset and compute the mean and median.

5. Alternatively, we can visualize and work with a *transformation* of our dataset. This is especially useful when we have a "long-tailed" dataset. Explain what the next few lines of code do.

```
ceo_salaries = ceo_salaries %>%
  mutate(log_pay = log(Pay + 1))

ceo_salaries %>%
  summarize(mean = mean(log_pay),
            median = median(log_pay))
```

```
# A tibble: 1 x 2
   mean median
  <dbl>  <dbl>
1  16.4   16.6
```

6. Create a histogram of the `log_pay` variable and describe the shape of the distribution.

7. Your answer to 6 should note that there's now *low* outlier(s)! Find these outlier(s). Which companies are they? What do they pay their CEOs? (*Note:* you should do this using code, but can check your answer with the "spreadsheet view" of the data.)

When you're done, please knit and submit the PDF to gradescope. Group submissions are enabled and you only need to submit one PDF per group (but please only include group members who were present).